IN THE SPECIFICATION

Please amend the specification as follows:

[0018] The invention is a method of representing biopolymers in a computational trainable system for use in the prediction of the interaction of proteins with other proteins, nucleic acids, small molecules and biopolymers. The interactions are determined in a pairwise fashion, with higher order structures containing more than two components being determined in multiple rounds of analysis. A collection of known biomolecular interactions, such as protein-protein interactions, are encoded as a set of features on a residue-by-residue basis in the trainable system. Databases of heterogenous protein-protein interactions exist, including the publicly-accessible Database of Interacting Proteins ~~(DIP: http://dip.doe-mbi.ucla.edu)~~ which at the time of this application contains 10933 interaction pairs. Other databases contain information regarding protein interactions in single organisms~~; one such database is available at http://pim.hybrigenics.com~~,which contains all of the known protein-protein interactions known in the bacteria *H. pylori*. The selection of a database is not a limiting aspect of the invention. Moreover, the databases listed should not be considered static entities or to be limited to the data that they contain at the time of the application. The databases are a source of training sets to "teach" the trainable system, but are not a component of the invention itself. The invention is instead the manner in which the biopolymers are represented as a linear set of features and used in the trainable system to predict the interactions of the encoded biopolymers with other molecules.

[0027] The invention is a method for the use of a trainable system to predict the presence of epitopes of interest, including functional domains and binding sites of proteins, and antigenic determinants. By casting the numerical optimization procedure as a regression problem, a continuous value for binding affinity of ligand-molecular complex can be learned. In this manner the same scheme for representing linear biopolymer sequences as features is used, and the training procedure involves "sliding" a window along the query sequence, each step outputting a numerical value that

constitutes a predicted interaction value of the sequence within the window and the query ligand. Example public-domain databases containing data appropriate for training the system in this mode are: (1) The Ligand Chemical Database for Enzyme Reactions (~~http://www.genome.ad.jp/dbget/ligand.html~~), (2) The Function Immunology Database of MHC molecules, antigens and diseases (~~FIMM; http://sdmc.krdl.org.sg:8080/fimm/~~), and (3) the ImMunoGeneTics database (~~IMGT; http://www.ebi.ac.uk/imgt/~~).

[0028] The invention is a method for the use of a trainable system to predict the binding of nucleic acids with proteins. This mode of prediction is carried out similarly to the antigenic determinant prediction scheme outlined above. Training data for local interactions between nucleic acid molecules (DNA, or RNA) and proteins are developed from the nucleic acid-protein complex structural data of the Protein Data Bank (~~PDB; http://www.rcsb.org/pdb/~~) and summarized in the DNA-Protein Interaction Database (~~DNAPIDB; http://www.dpidb.belozersky.msu. ru/~~). The sites of interaction are analyzed as before and converted to a set of features in the learning machine. The trained system outputs a thresholded-score indicative of the local propensity for nucleic acid binding at each site along the query protein.

[0030] The invention is a method for cell-map proteomics. Biochemical, signaling and gene regulatory path ways can be mapped for entire organisms. The entire genome of the *Helicobacter pylori*, which contains coding sequences for 486 proteins, has been sequenced and 1,039 protein-protein interactions have been mapped. Using this model organism, which performs all of the functions required for viability, one can map the interactions of genomes of similar organisms, such as *Campylobacter jejuni*, an enteric bacteria pathogen that causes common symptoms of food poisoning. ~~A complete protein-protein interaction map for *C. jejuni* computed using the methods disclosed herein is available at http://www-bioeng.ucsd.edu/cjbean/.~~ Analysis of the major constituent protein domains shows a high degree of similarity. These orthologous bacterial proteomes represent a model system for demonstrating the utility of the invention for performing proteome wide interaction mining. The accuracy of

the proteome map will depend on the quality of the database as well as the level of similarity of the organisms to be analyzed. The higher the similarity and the greater the number of interactions defined, the greater the predictive value of the information in the database.

[0031] *Databases of known biomolecular interactions.* Databases of protein interactions are available at multiple sites including the Database of Interacting Proteins (DIP) ~~http://dip.doe-mbi.ucla.edu~~ which currently contains 10933 entries, and the *H. pylori* database, http://pim.hybrigenics.com which contains 1273 interacting pairs between the 486 potential proteins of the organism. In the DIP database, each interaction pair contains fields representing accession codes for other pubic protein databases, protein name identification and references to experimental literature underlying the interacting residue ranges, and protein-protein complex dissociation constants. The protein interaction domain coverage within the DIP is diverse; at least 175 distinct domains are represented. The proteins are predominantly eukaryotic, with a majority of the proteins being from the yeast *Saccharomyces cerevisiae*. The information in the database is updated constantly by individuals studying protein-protein interactions, thus providing an increasing number of interactions that may be "taught" to the trainable system of the invention.

[0032] ~~A summary of public domain databases containing data appropriate for training this invention are listed in the following table. The entries in this table represent only a small subset of currently available databases, which continue to appear and grow in size.~~

PLEASE DELETE TABLE 1.

0041] Support Vector Machine learning was implemented using Joachims' SVM^{light} (Joachims, 1999). ~~, available online at http://www-ai.cs.uni-dortmund.de/SOFTWARE/SVM_LIGHT.eng.html.~~

[0047] Training data for local interactions between nucleic acid molecules (DNA, or RNA) and proteins can be developed from the nucleic acid-protein complex structural data of the Protein Data Bank ~~(PDB; http://www.rcsb.org/pdb/)~~ and summarized in the DNA-Protein Interaction Database ~~(DNAPIDB; http://www.dpidb.belozersky.msu.ru/)~~. The sites of interaction are analyzed as before and converted to a set of features in the learning machine. The trained system outputs a thresholded-score indicative of the local propensity for nucleic acid binding at each site along the query protein.

[0048] *Prediction of protein epitopes*. The invention is a method for the use of a learning machine to predict the presence of epitopes of interest, including functional domains and binding sites of proteins, and antigenic determinants. The learning algorithm in this application is cast as a regression similarly to the DNA/RNA-protein determinant prediction scheme outlined above._Example public-domain databases containing data appropriate for training the system in this mode are: (1) The Ligand Chemical Database for Enzyme Reactions ~~(http://www.genome.ad.jp/dbget/ligand.html)~~, (2) The Function Immunology Database of MHC molecules, antigens and diseases ~~(FIMM; http://sdmc.krdl.org.sg:8080/fimm/)~~ , and (3) the ImMunoGeneTics database ~~(IMGT; http://www.ebi.ac.uk/imgt/)~~ .